Preservation Watch: um sistema de suporte à preservação digital

José Carlos Ramalho jcr@keep.pt
KEEP SOLUTIONS www.keep.pt

Luís Faria lfaria@keep.pt
KEEP SOLUTIONS www.keep.pt

Miguel Ferreira mferreira@keep.pt
KEEP SOLUTIONS www.keep.pt

Encontro Internacional de Arquivos
Évora, Portugal, 2014-10-03

KEEP SOLUTIONS: Projetos

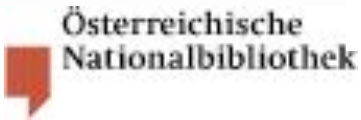
University of Minho SPIN-OFF

- DigitArq, CRAV (2003..[2008-2012])
- RODA (2006..[2008-…[)
- RCAAP (2008-…)
- PPA (2009)
- Open source: RODA, KOHA, DSpace, Moodle, etc.
- Scientific research
  - **SCAPE**: Preservação digital em larga escala
  - **4C**: previsão de custos na preservação digital
  - **e-arK**: desenvolvimento de um modelo de referência europeu baseado no OAIS

http://www.keep.pt

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

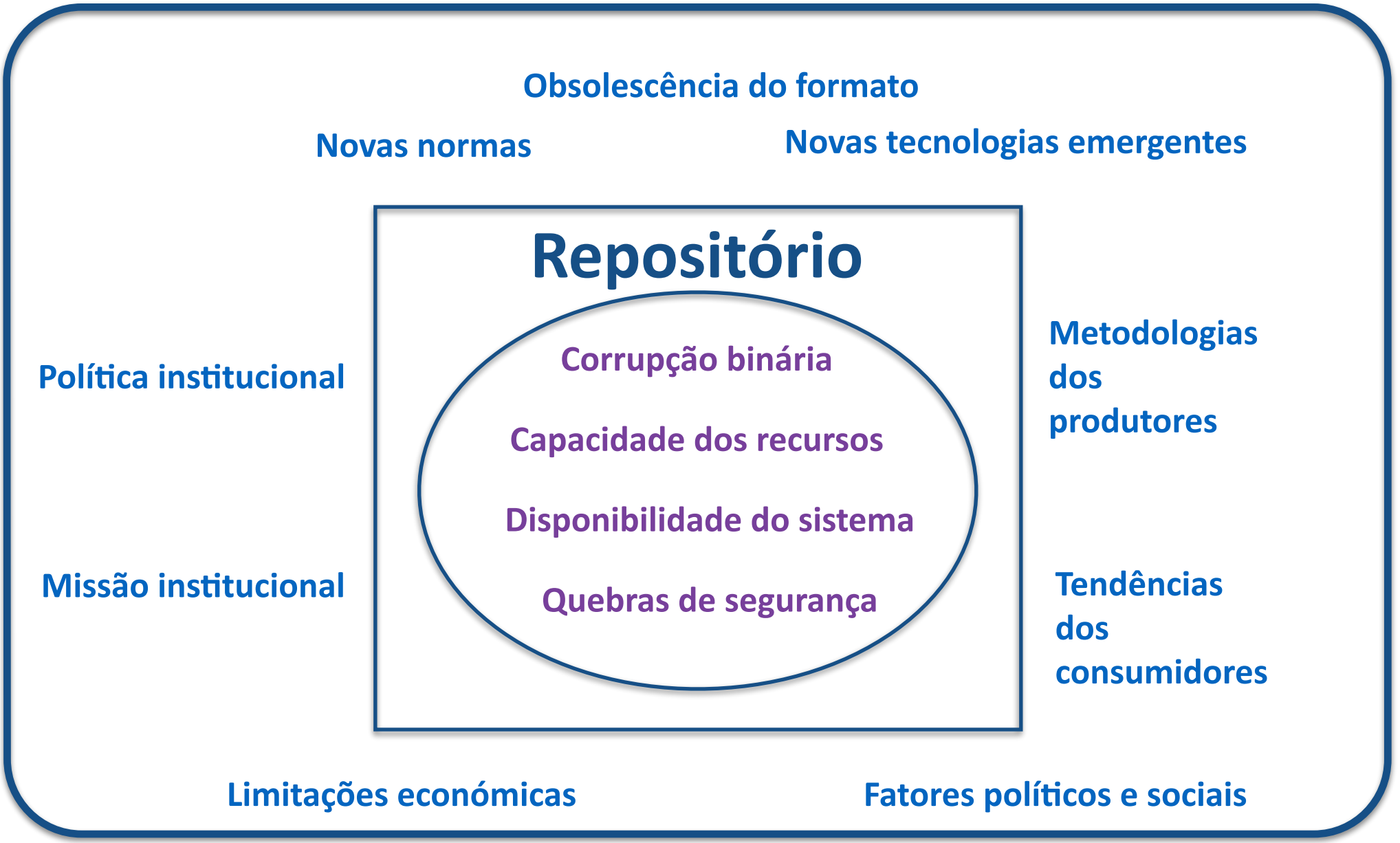
Parceiros



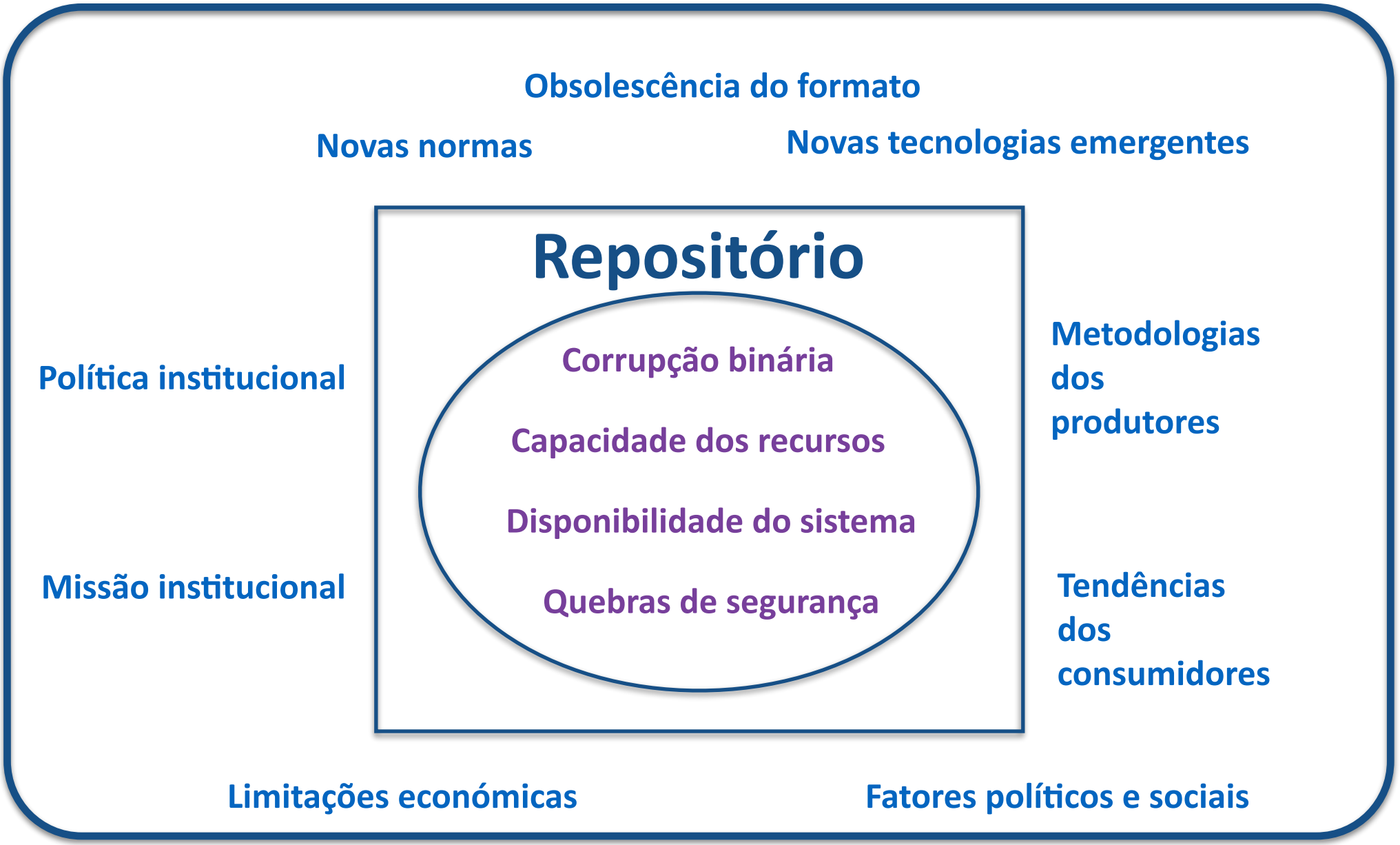
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Monitorização da Preservação Digital

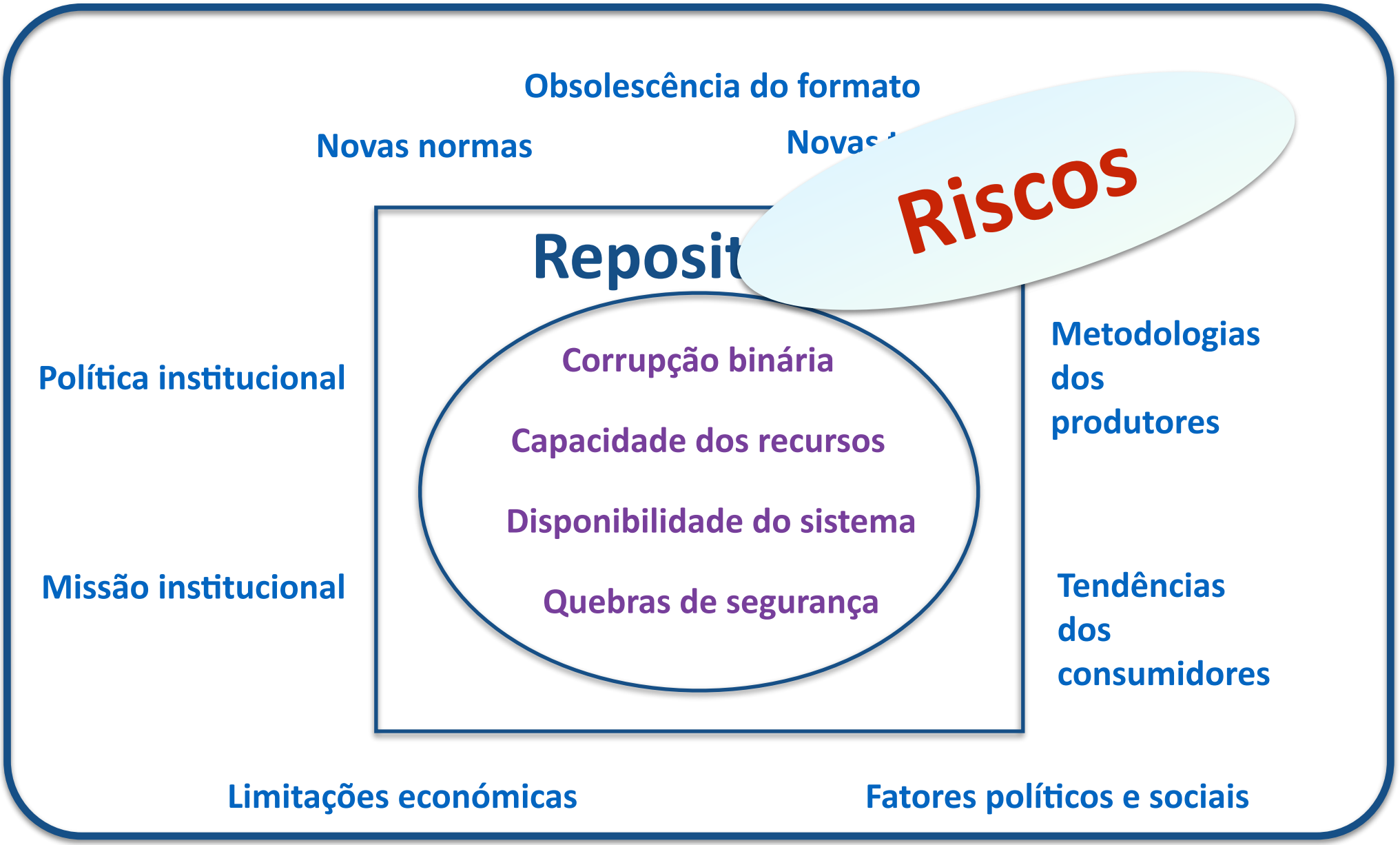
Porque necessitamos de monitorização?



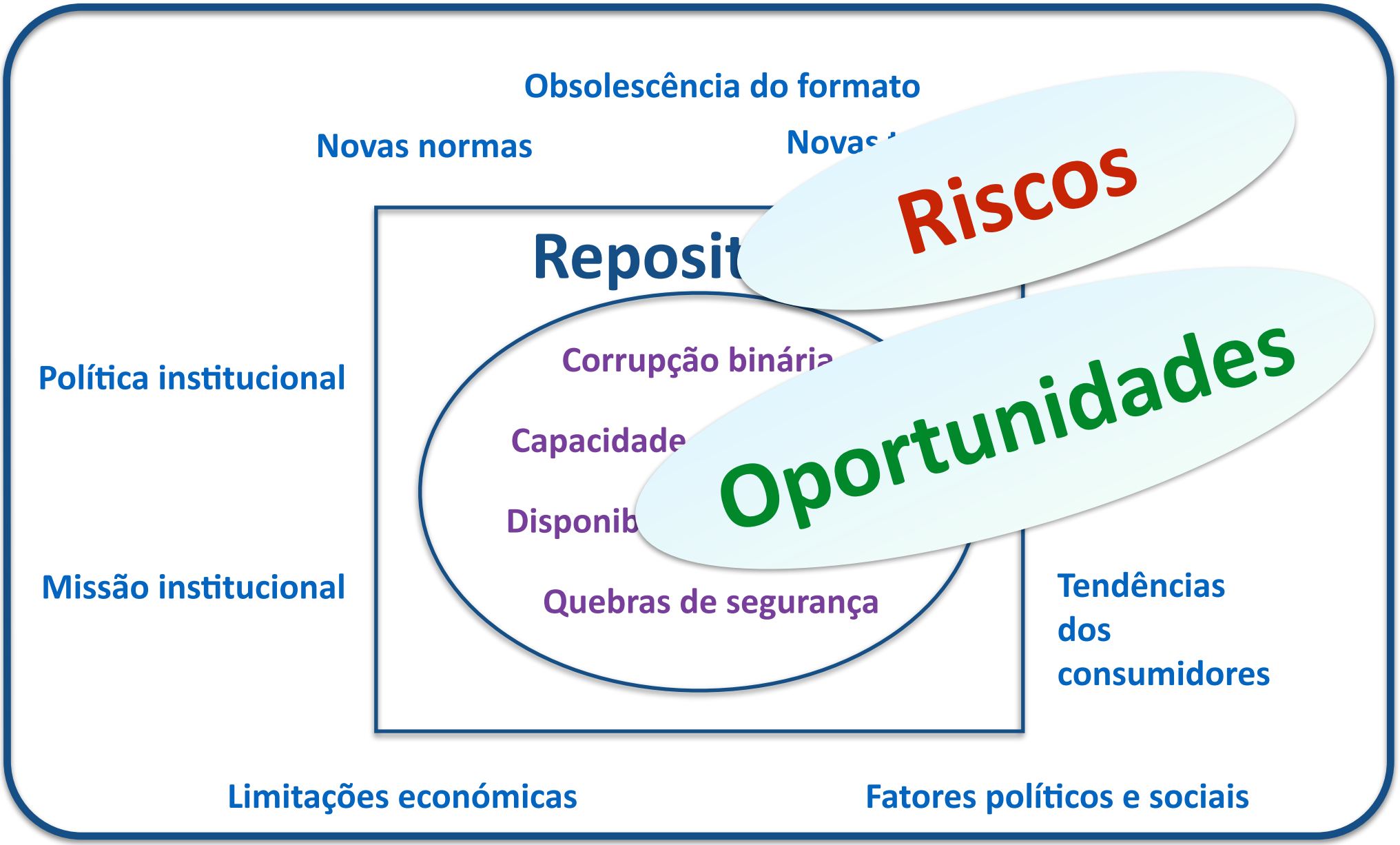
Porque necessitamos de monitorização?



Porque necessitamos de monitorização?



Porque necessitamos de monitorização?



Estado da Arte

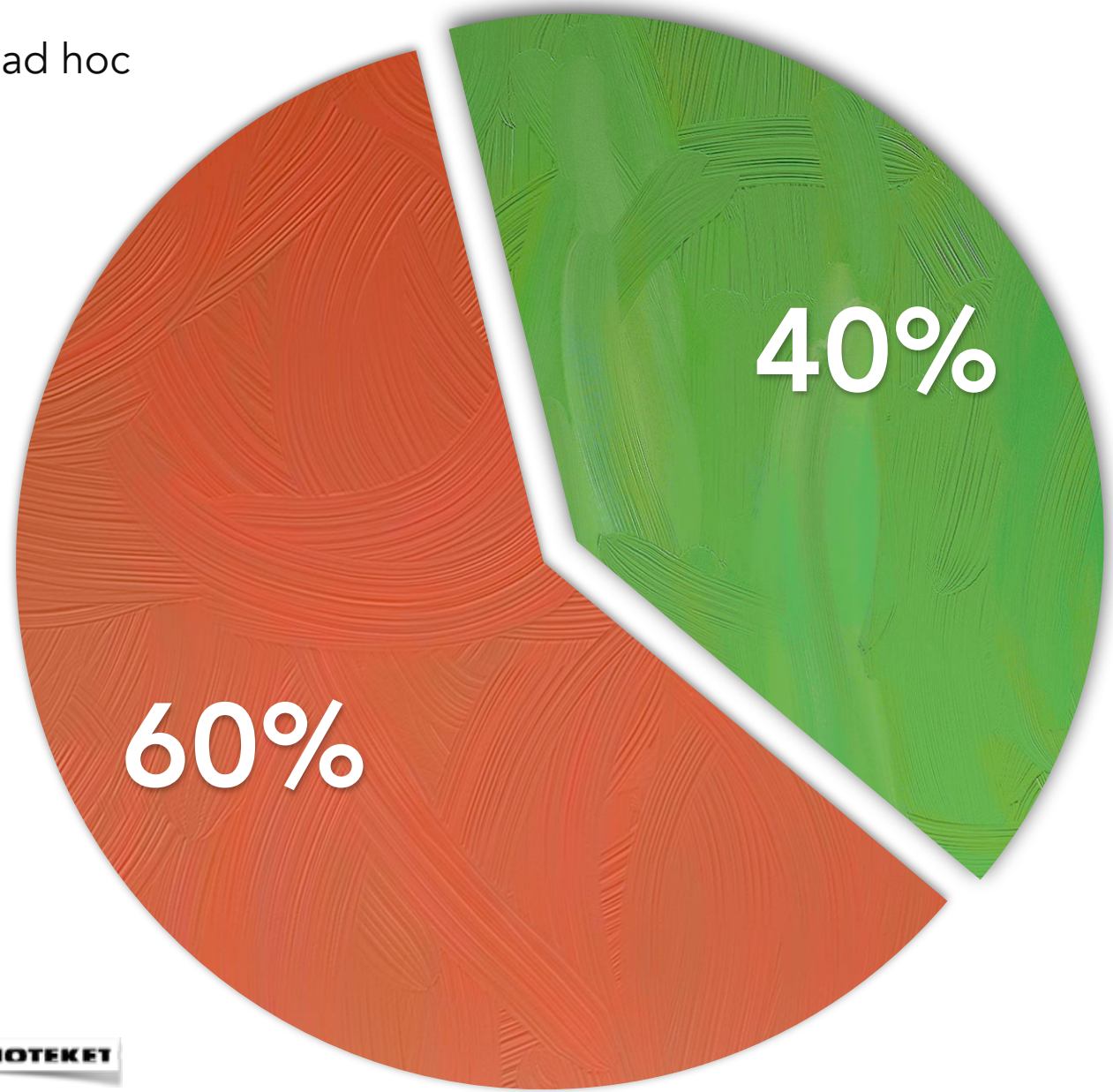
- Digital Format Registries
- Automatic Obsolescence Notification System (AONS)
- Relatórios de vigilância tecnológica

Estado da Arte

- Digital Format Registries
  - Falta de cobertura
  - Riscos genéricos definidos estaticamente
  - Riscos não estruturados
  - Focado na obsolescência do formato
- AONS
  - Totalmente dependente dos registos de formato
- Relatórios de vigilância tecnológica
  - Inacessíveis às máquinas (elegíveis)

Avaliação de Risco

- Sim, mas manualmente e ad hoc
- Não



Participantes:

Science & Technology Facilities Council

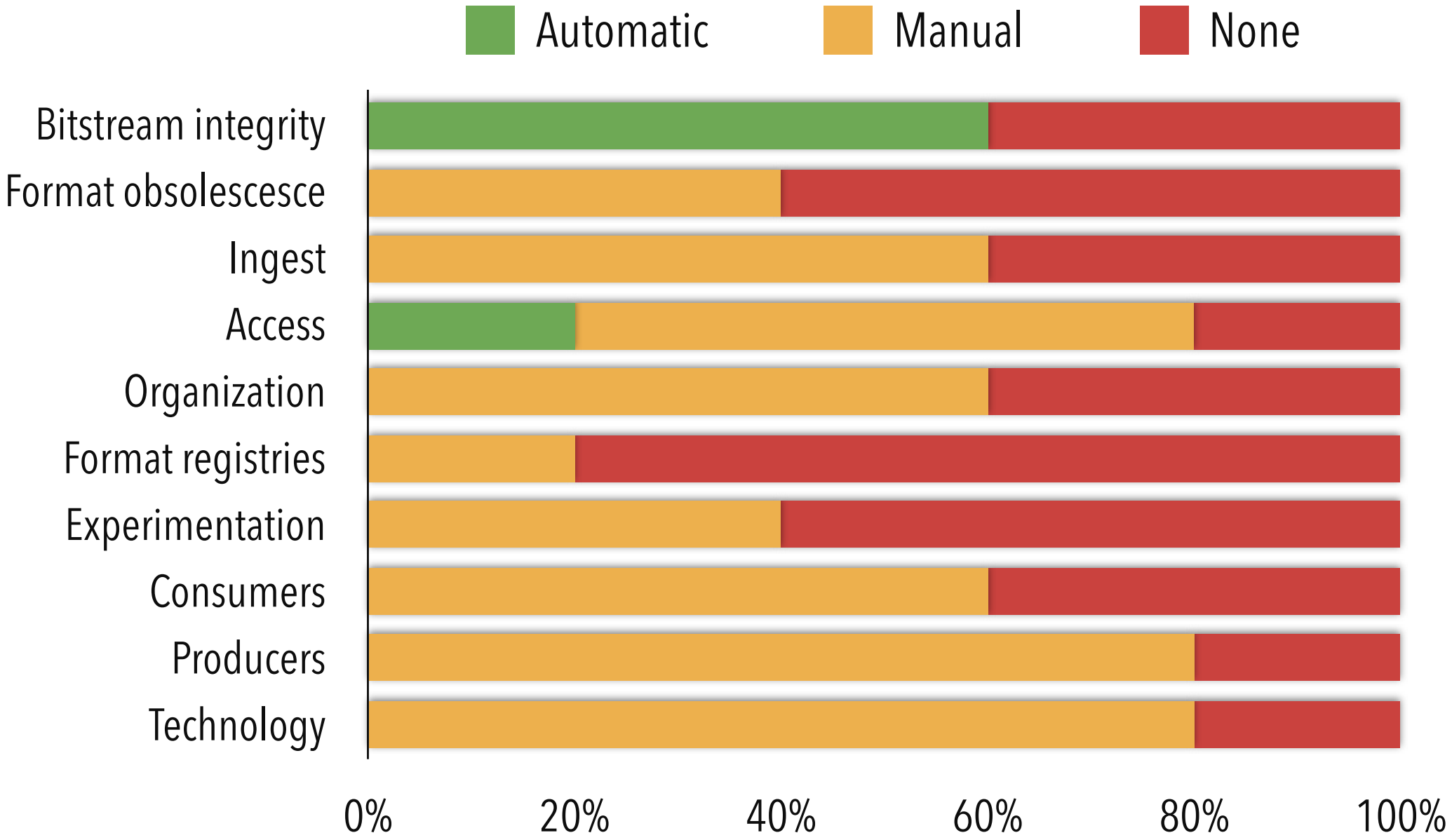
INTERNET MEMORY FOUNDATION

BRITISH LIBRARY

STATSBIBLIOTEKET

KB Koninklijke Bibliotheek National Library of the Netherlands

Monitorização

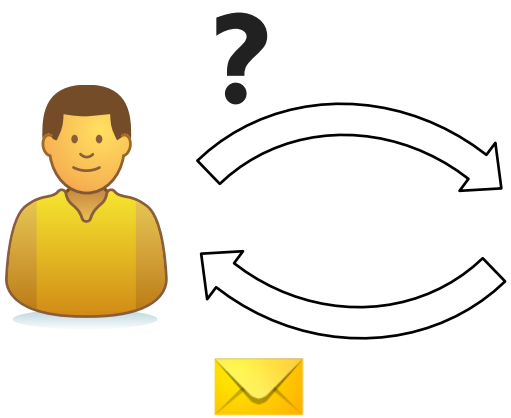


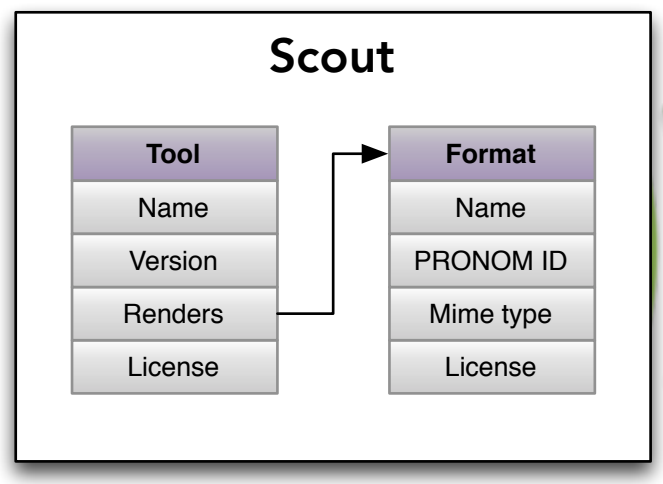
O que é necessário?

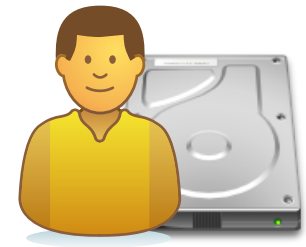
- Precisamos de informação!
  - De todo o lado e de toda a gente
  - Partilhando
- Escalabilidade e usabilidade
  - Dados estruturados
  - Vocabulário controlado

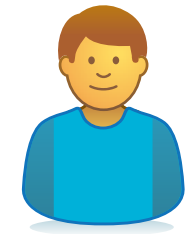
Scout

Uma nova aproximação









ConversionSoftwareRegistry

PRONOM

UDFR

FILExt
A free online resource by Uniblue

FileInfo.com

alternativeTo

i use this

CNET | Download.com

Objetivos

- Coletar informação de várias fontes
- Permitir a introdução manual de dados
- Base de dados centralizada para suporte à preservação digital
- Permitir que os utilizadores coloquem questões
- Notificar os utilizadores quando ocorrem mudanças ou eventos significativos

Problemas

- Um Repositório alberga conteúdos
- Uma Organização tem políticas em curso (e.g. não são permitidos conteúdos comprimidos)

P1: Será que os conteúdos respeitam as políticas vigentes? Há algum risco associado?
Mesmo que conteúdo, política e ambiente estejam em constante mudança?

- Encontramos um risco na preservação digital!

P2: Como decidiremos a ação a tomar mantendo os requisitos de confiança e autenticidade?

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Mais problemas

- Saber que ação tomar

P3: Como monitorizar a qualidade da ação tomada e como garantir que os invariantes de preservação se mantêm?

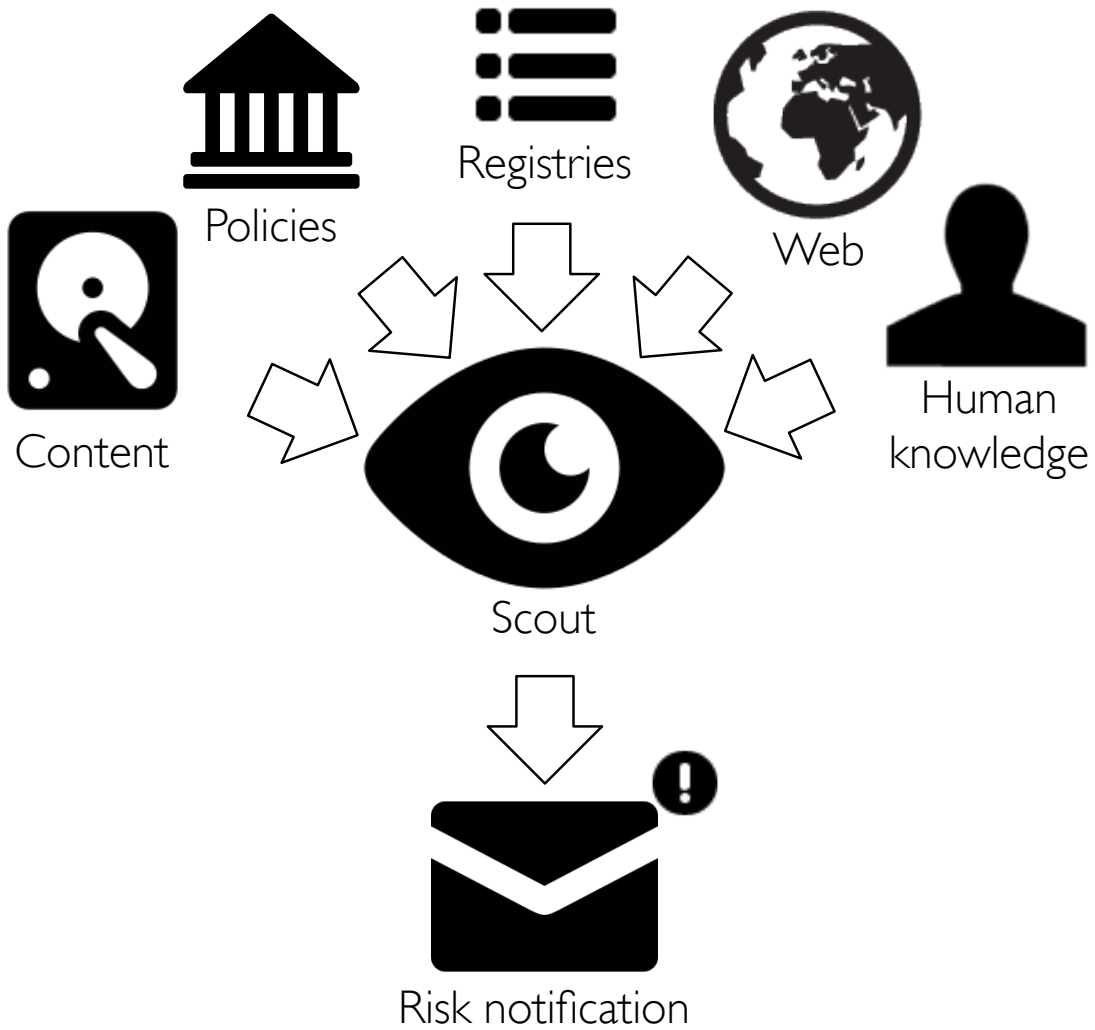
- Os conteúdos crescem exponencialmente em volume, heterogeneidade e complexidade

P4: Como implementar a preservação digital em sistemas de grande escala (big data)?

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Scout: a preservation watch system

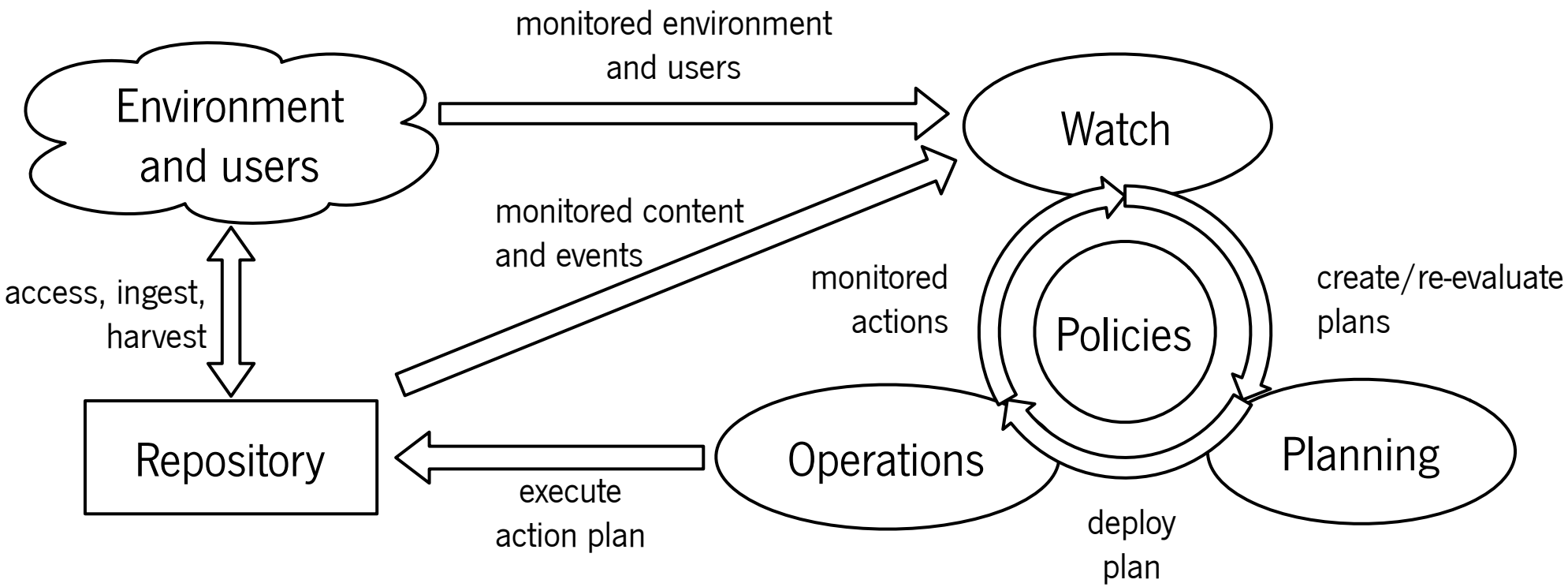
- Monitoriza facetas do mundo para detetar riscos e oportunidades de preservação
- Triple store
- Interoperabilidade
  - Data Connector & Report API
  - SCAPE Policy model
  - PRONOM
  - Web semantic extraction
  - Renderability experiments
- Interface Web
- Alertas: templates e SPARQL
- Notificações por email
- Demo: http://scout.scape.keep.pt



http://openplanets.github.io/scout/

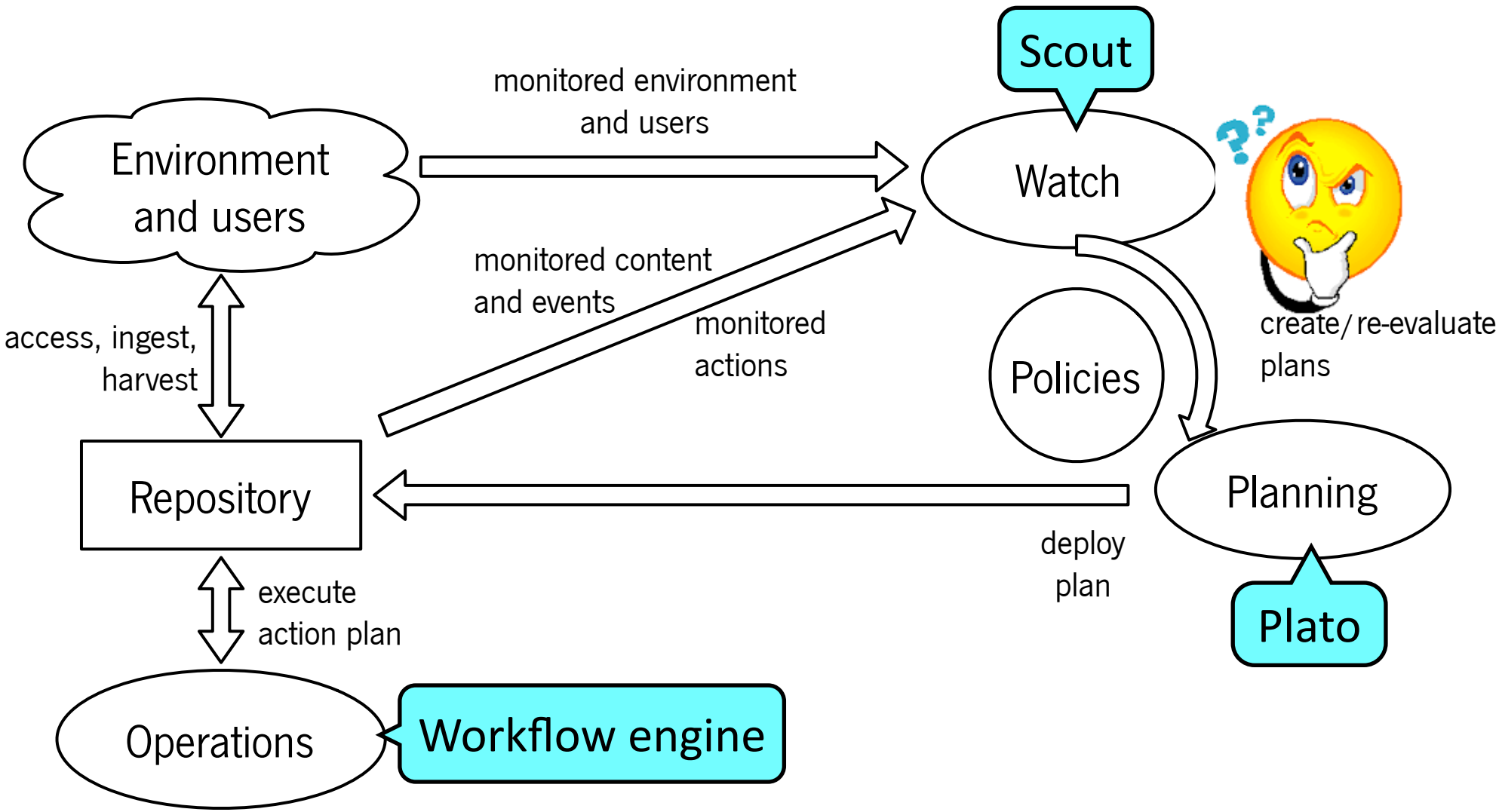
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Ciclo de vida da preservação: cenário ideal



This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Ciclo de vida da preservação: na prática



This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

API para transferência de dados (interoperabilidade)

- Permite aceder e modificar conteúdos no repositório
- HTTP REST API
- Methods:
  - **Retrieve** entidade inteletual, metadados, representação, ficheiro ou bit stream
  - **Ingest** entidade inteletual (sync ou async)
  - **Update** entidade inteletual, representação ou ficheiro
  - **Search** entidades, representações ou ficheiros (SRU)
- Especificação da API: https://github.com/openplanets/scape-platform-api
- Implementação de ref.: Fins de 2013 no Fedora 4 e no RODA

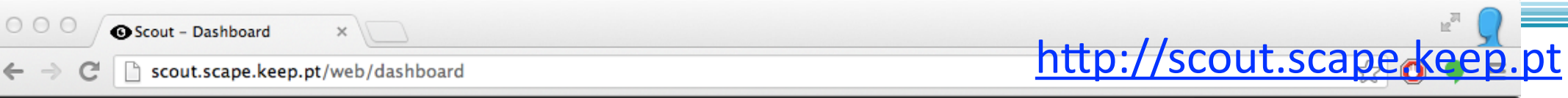
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

API de relato

- Dá acesso aos eventos do repositório
- Eventos:
  - **Ingestão:** início e fim
  - **Visualização** ou **descarga:** metadados descritivos ou representações
  - **Execução** de planos de preservação
- Fornecedor OAI-PMH
- Metadados PREMIS associados aos eventos
  - Agent: **quem** acionou o evento
  - Date/time: **quando** é que o evento ocorreu
  - Details: **que** aconteceu
- API: https://github.com/openplanets/scape-platform-api
- Implementação de ref.: https://github.com/openplanets/roda

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

http://scout.scape.keep.pt



Dashboard

All about your own information.

My triggers

You have no triggers defined, create one now!

+ Create trigger

My policies

Objective	Measure	Description	Modality	Qualifier	Value
0	Running costs per object	Running operational costs of an action in € per object.	MUST	LT	0.24
1	elapsed time per MB	elapsed processing time per Megabyte of input data, measured in milliseconds	MUST	LT	2000
2	stability judgement	Judgement of the stability of an action	SHOULD		stable
3	ease of integration	Assessment of how easy it is to integrate an action into a particular server environment.	SHOULD		good
4	software licence source code	Indicates if and in which way the source code of the software is accessible.	MUST		openSource
5	ease of use	Assessment of how easy it is to use an action in operations	SHOULD		openSource
6	image width equal	true iff image width has been preserved.	MUST		true

Collection size The overall size

http://scout.scape.keep.pt

43.97 GB

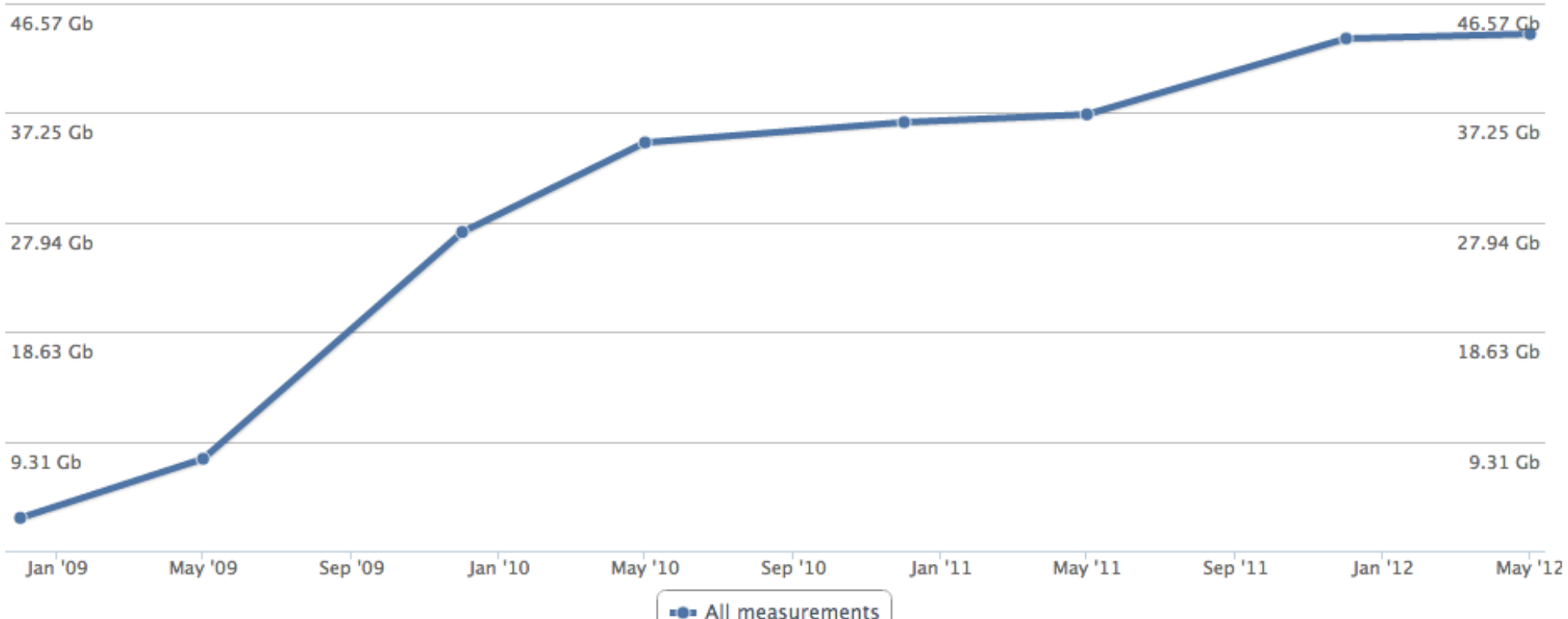
Data type: Very big integer number (File or storage size).

Property history: There are 8 different values of this property, this is number 7 (starts at 0).

Value provenance: Current value was measured 1 times by 1 different sources.

Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.



Format distribution The Format distribution of the objects

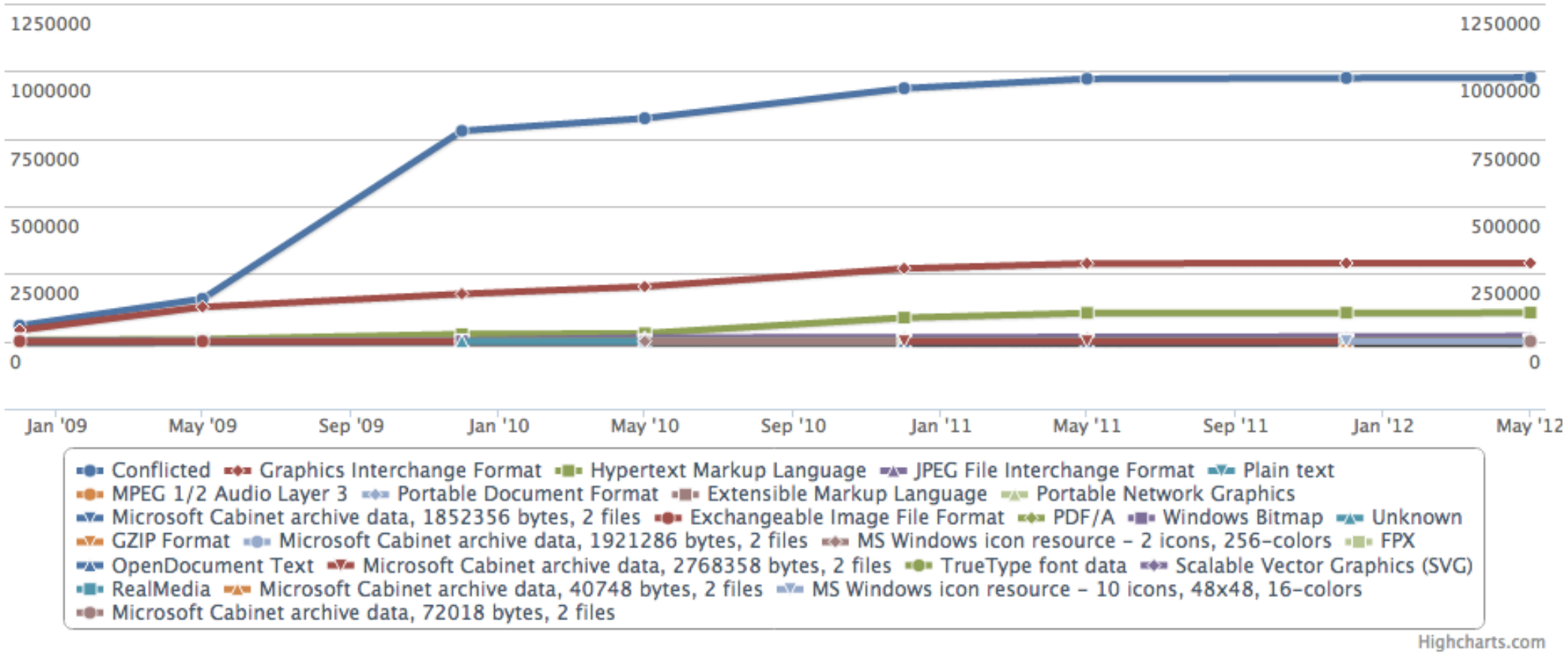
http://scout.scape.keep.pt

Key	Value
Tagged Image File Format	160
Hypertext Markup Language	23
Portable Document Format	17
Plain text	16
XLS	16
FPX	9
Microsoft Word	7
Extensible Markup Language	2
Extensible Hypertext Markup Language	2
Postscript	2
Macromedia Flash data (compressed), version 6	1
Macromedia Flash data, version 5	1
PPT	1
news or mail, ASCII text	1

Property history

http://scout.scape.keep.pt

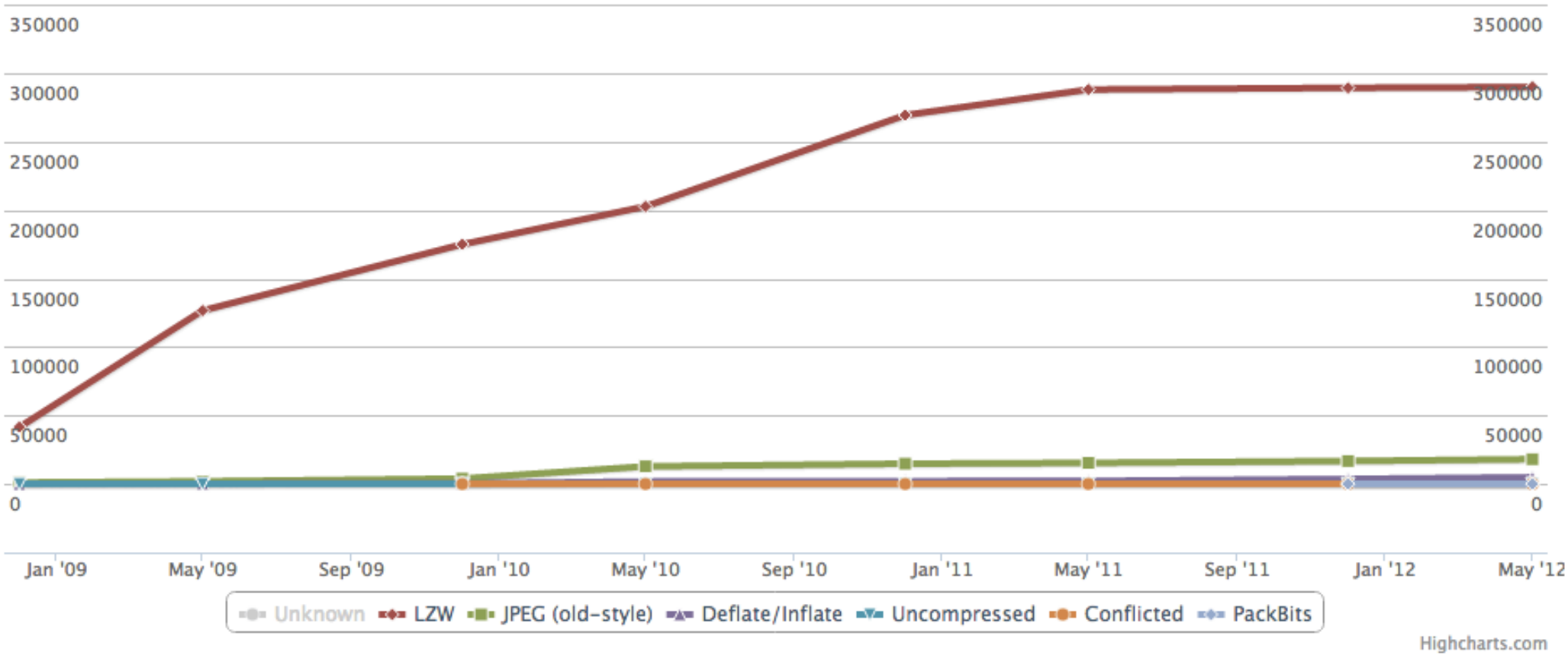
This property has changed in time as represented in the chart below. Click on the chart dots for more information.

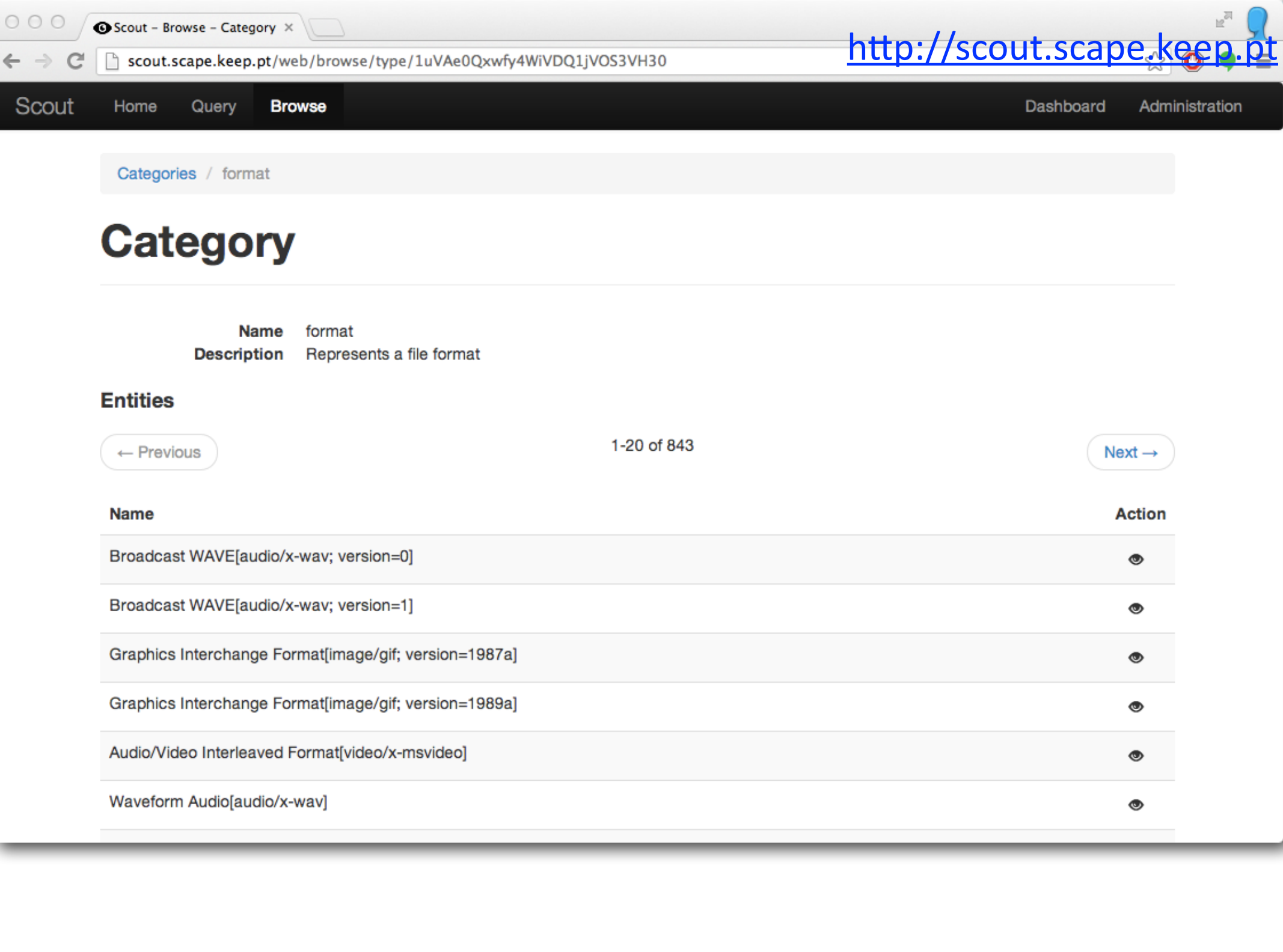


Property history

http://scout.scape.keep.pt

This property has changed in time as represented in the chart below. Click on the chart dots for more information.




Scout - Browse - Category
scout.scape.keep.pt/web/browse/type/1uVAe0Qxwfy4WiVDQ1jVOS3VH30
http://scout.scape.keep.pt
Scout
Home
Query
Browse
Dashboard
Administration
Categories / format
Category
Name format
Description Represents a file format
Entities
← Previous
1-20 of 843
Next →
Name
Action
Broadcast WAVE[audio/x-wav; version=0]
Broadcast WAVE[audio/x-wav; version=1]
Graphics Interchange Format[image/gif; version=1987a]
Graphics Interchange Format[image/gif; version=1989a]
Audio/Video Interleaved Format[video/x-msvideo]
Waveform Audio[audio/x-wav]

Properties

http://scout.scape.keep.pt

Name	Value	Action
Minimum preservation action execution time	1.5002512	
Average preservation action execution time	1.8746954	
Maximum preservation action execution time	2.3340003	
Ingest average time (ms)	1092798.0	

Advanced query

http://scout.scape.keep.pt

Use SPARQL to make your own query

Target

- (•) Category
- ( ) Property
- ( ) Entity
- ( ) Value
- ( ) Measurement

Snippets

Relations

Resources

SPARQL Help

```
SELECT ?s WHERE { ?s rdf:type watch:EntityType .

}
```

+ Create trigger

Search

Query

http://scout.scape.keep.pt

Select a pre-made question template or go to advanced query.

Query templates

- **Check collection policy conformance**
- Collection size limit

Check collection policy conformance

Check if selected collection conforms to the defined policy (only compression scheme policy is checked right now)

Collection: The ID from the URL

Your collection profile already inserted into scout

Search | + Create trigger

Conclusions

P1: Será que os conteúdos respeitam as políticas vigentes? Há algum risco associado?
Mesmo que conteúdo, política e ambiente estejam em constante mudança?

S1: Utilize o Scout: preservation watch system

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Conclusions

P2: Como decidiremos a ação a tomar mantendo os requisitos de confiança e autenticidade?

S2: Utilize o Plato: preservation planning tool

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Conclusions

P3: Como monitorizar a qualidade da ação tomada e como garantir que os invariantes de preservação se mantêm?

S3: Q&A in preservation plans (Plato), monitoring of Q&A (Report API & Scout), automatic Scout triggers created by Plato

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Conclusions

P4: Como implementar a preservação digital em sistemas de grande escala (big data)?

S4: Automação e integração dos processos de preservação.

Caso de estudo do SCAPE: FITS + C3PO

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Roadmap

- Scout:
  - Suporte de utilizadores
  - Mais conetores
  - Mais templates para alertas
- Plato:
  - Criação automática de alertas no Scout
  - Publicação automática usando a API de gestão
- Implementações de um Repositório de referência: RODA e Fedora 4

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

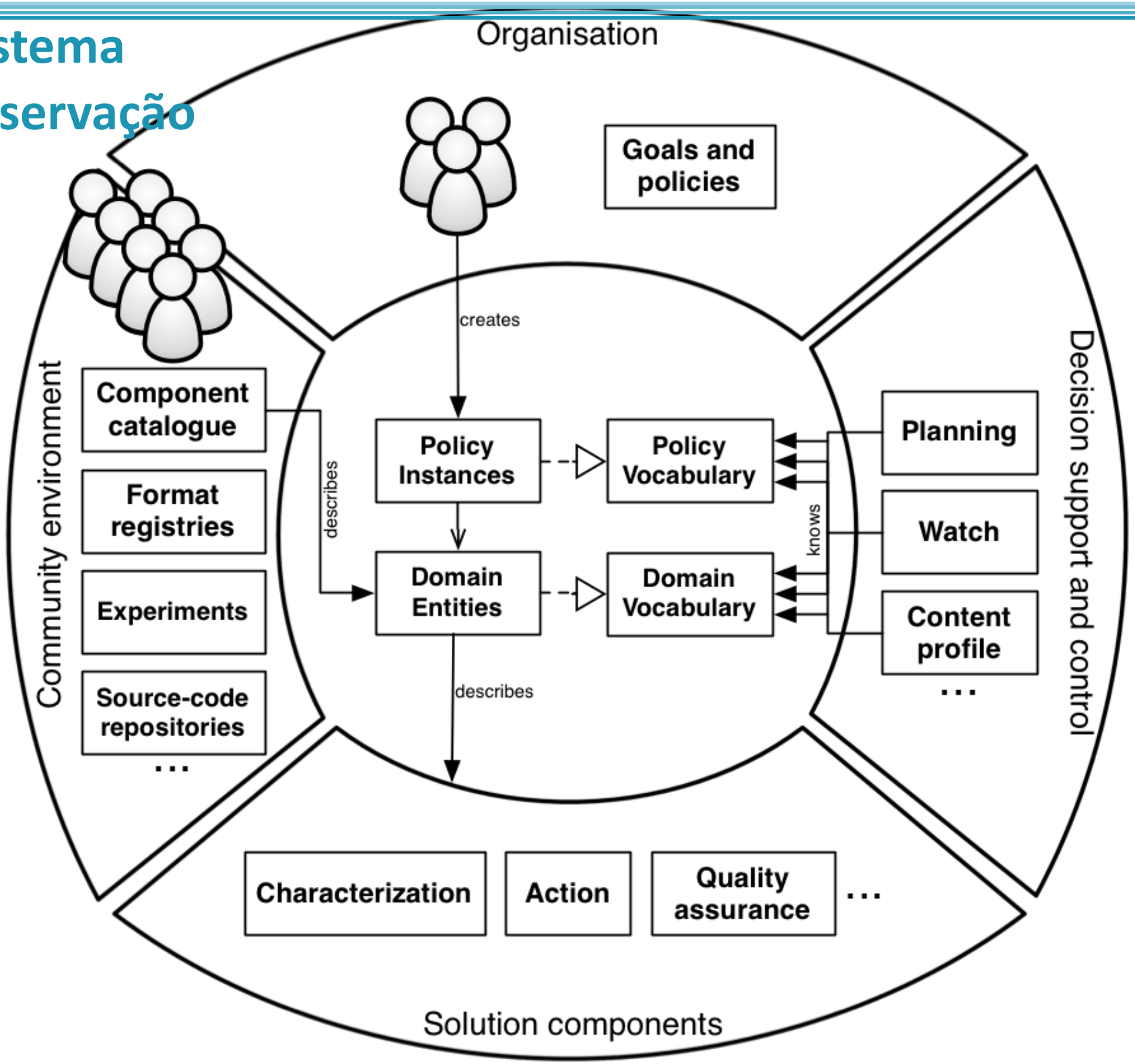
Conclusions

- Todas as APIs estão publicadas
- Implementações de referência: RODA e Fedora 4
- Todas as ferramentas disponíveis no Github

Adiciona uma política de preservação ao teu repositório já!

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Ecossistema de Preservação



História dum ciclo de vida de Preservação (parte 1)

- Um Repositório tem conteúdos
- A Organização tem políticas em vigor (e.g. não permitir compressão)
- Formaliza as políticas
- Usa o Scout para monitorizar a conformidade
  - Carrega as políticas no Scout
  - Cria adaptadores para o teu repositório
  - Cria alertas
  - Recebe notificações: há ficheiros comprimidos!
- Usa o C3PO para analisar em detalhe o problema
  - Podes ter que dividi-lo em problemas mais pequenos

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

História dum ciclo de vida de Preservação (parte 2)

- Usa o Plato para encontrar uma solução para o problema:
  - Carrega as políticas: objetivos automáticos
  - Encontra ferramentas alternativas automaticamente
  - Testa as ferramentas automaticamente com amostras de conteúdo
  - Encontra a melhor alternativa
  - Cria um plano automaticamente com documentação, ações e Q&A
  - Envia o plano diretamente ao repositório e os alertas ao Scout
- Executa o plano no motor de workflow
  - O Repositório executa o plano diretamente no motor de workflow
  - Os resultados são agregados através da API do conetor de dados
  - As ações de preservação e as Q&A são enviadas ao Scout via a API de relato
- O Scout deteta os riscos que têm de ser resolvidos

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Questões?

José Carlos Ramalho

Consultor / Investigador

jcr@keep.pt / jcr@di.uminho.pt

KEEPSOLUTIONS

University of Minho SPIN-OFF

ARQUIVOS | BIBLIOTECAS | MUSEUS

www.keep.pt